Supplementary file

Sex difference of mutation clonality in diffuse glioma evolution

Hongyi Zhang[1,†], Jianlong Liao[1,†], Xinxin Zhang[1,†], Erjie Zhao[1,†], Xin Liang[1], Shangyi Luo[1], Jian Shi[1], Fulong Yu[1], Jinyuan Xu[1], Weitao Shen[1], Yixue Li[1], Yun Xiao[1,*], Xia Li[1,*]

1 College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang 150081, China.

† These authors contributed equally to this work.

* To whom correspondence should be addressed

Email: lixia@hrbmu.edu.cn, xiaoyun@ems.hrbmu.edu.cn

Supplementary Methods

Power for detecting low-frequency mutations

We adopted a previously described framework to calculate the statistical power for detection of somatic single nucleotide variants (sSNV)[1]. This power is depended on the theoretic allele fraction *f* and local sequencing coverage *n* of a variant. Assuming the random sequencing error is *e* ($e = 1 \times 10^{-3}$),the probability of observing at least *m* identical alternate reads due to sequencing error can be represented as:

$$P(m) = \begin{cases} 1 & if \ \ m = 0 \\ 1 - \sum_{i=0}^{m-1} Binom(i \mid n, e/3) & if \ \ m \geq 0 \end{cases}$$

The minimum number of alternate reads *k* supporting that the *P(k)* is less than a given false-positive rate (FPR) is:

$$k = \min(\mathrm{m}) \mid P(m) \leq \mathrm{FPR}$$

Let FPR=5×10^{-7},we obtained those variants with *k* or more alternate reads, which were considered to be detected. Then, the detection power of each sSNV is calculated as:

$$Pow(n, f) = 1 - \sum_{i=0}^{k-1} Binom(i \mid n, f) + \frac{Binom(k \mid n, f)\left(FPR - P(k)\right)}{P(k-1) - p(k)}$$

For a clonal sSNV, the theoretic allele fraction *f* is denoted as f_c:

$$f_c = \frac{p}{2(1-p) + p * CPN_{mut}}$$

Therefore, the detection power of the clonal sSNV is given by $Pow(n,f_c)$. While the power for subclonal variants is calculated as $Pow(n,sf_c)$, where *s* represents the estimated CCF of subclonal mutations. In this study, we retained the point mutations with detection power >80% and all indel mutations for subsequent analysis.

Adjusting confounders

We performed confounder adjustments based on propensity scoreto reveal the association between clonal/subclonal mutation load and gender[2]. The patient characteristics including age at diagnosis, tumor purity, race, *IDH1*/*IDH2* mutation(G-CIMP status), 1p/19q co-deletion, WHO grade and histologic subtype were used as covariates. We first calculated the propensity score based on "sex" using logistic regression. Then we employed the nearest available matching scheme on the estimated propensity score[3]. This step removed 118 male samples (GBM: 69, LGG: 49), which balanced the propensity scores and further the covariates/confounders (Table S6 and S7). The standardized differences (SD) were used to check balance for each covariate and propensity score:

$$SD = \frac{\overline{x_{fe}} - \overline{x_{ma}}}{\sqrt{(s_{fe}^2 + s_{ma}^2)/2}}$$

Where for each covaraiate $\overline{x_{fe}}$ and $\overline{x_{ma}}$ are the sample means in the female and male groups, respectively, and s_{fe}^2 and s_{ma}^2 are the corresponding sample variances. The covariates with |SD| < 10% were considered as balanced[4]. Using balanced data, we could unbiasedly compare clonal and subcloanl mutation burden between two sex groups of GBM and LGG.

Validation of the prognostic value of *PTEN* clonality in other GBM cohort

To validate the association between subclonal mutation in *PTEN* and poor survival of GBM female patients, we obtained another GBM cohort from cBioPortal, which contained 268 primary samples provided by MSKCC (memorial sloan-kettering cancer center). These samples were measured on capture–based sequencing in 341 genes including *PTEN*[5]. We focused on GBM samples without copy number alterations of *PTEN*, so that we could let the local copy number of *PTEN* be equal to 2 (CPN_{mut} = 2).The tumor purity *p*of each sample was annotated by a molecular pathology fellow[5]. Next, we used the same framework described in main text to infer the cancer cell fraction (CCF) and clonality of somatic point mutations in *PTEN*. Totally, 49 clonal and 13 subclonal mutations of *PTEN* were identified and Kaplan–Meier survival analysis validated the prognostic value of *PTEN* clonality in GBM females (p = 0.046, Figure S9), but not in males (p = 0.91). Multivariate cox regression analysis was performed with additional covariates including tumor purity, GCIMP status and smoking history. We observed significant correlation between clonal status of *PTEN* and overall survival of GBM female patients (p = 0.0332, HR = 23.7, 95% CI = 1.28 to 435.2) independent of tumor purity, GCIMP status and smoking history.

Comparing the clonal fraction of *IDH* mutations between TCGA and other data

When analyzing TCGA data, we observed more subclonal *IDH* mutations. However, *IDH1/2* mutations have been proposed as early clonal events during gliomagenesis.

Therefore, to exclude the influence of the method, we obtained the data of Zehir et al.[5]. To estimate the clonal fraction of *IDH* mutations in LGG. Zehir et al. implemented a hybridization capture–based sequencing in 341 cancer-associated genes including *IDH1* and *IDH2*. Their data includes the sequence data from 216 LGG patients, of which the tumor purities of 207 primary samples were also provided, enabling us to estimate the clonal status of *IDH* mutations. We used the same method and parameters to infer the clonal status of *IDH1/2* mutations. Among the 149 *IDH*-mutant cases, we identified 127 clonal (*IDH1*: 116, *IDH2*: 11) and 22 subclonal (*IDH1*: 22, *IDH2*: 0) mutations of *IDH*. The clonal fraction of *IDH* mutations reaches 85% (127/149), consistent with the previous viewpoint that *IDH* mutation tend to be an early event in LGG, suggesting the reliability of the method.

We additionally obtained the mutation data of primary LGG samples collected by Johnson et al.[6]. Johnson et al. sequenced the exomes of initial low-grade gliomas and recurrent tumors resected from 23 patients, and 30 *IDH*-mutated samples of 23 primary tumors (some different samples were derived from a same case) were identified. We compared the VAFs of *IDH* mutations between this data and TCGA data, and observed that the VAFs of *IDH* mutations in LGG samples used by Johnson et al. were significantly higher than those of TCGA LGG samples (median: 0.43 vs 0.35, p=0.00036, Wilcoxon rank-sum test). Assuming that the purities of tumor samples in these two data sets follow similar distributions, this result supports the lower clonal fraction of *IDH* mutations observed in TCGA LGG samples.

References

1. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol.* May 2012;30(5):413-421.

2. Rosenbaum PR, Rubin, D.B. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat.* 1985(39):33-38.

3. D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine.* Oct 15 1998;17(19):2265-2281.

4. Yuan Y, Liu L, Chen H, et al. Comprehensive Characterization of Molecular Differences in Cancer between Male and Female Patients. *Cancer cell.* May 9 2016;29(5):711-722.

5. Zehir A, Benayed R, Shah RH, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature medicine.* Jun 2017;23(6):703-713.

6. Johnson BE, Mazor T, Hong C, et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science.* Jan 10 2014;343(6167):189-193.

Supplementary Table Legends

Table S1 Detailed patient statistics across GBM and LGG cohorts

Table S2 The mutation drivers of GBM
Attachment

Table S3 The mutation drivers of LGG
Attachment

Table S4 Sex comparisons of other factors in raw GBM data

Table S5 Sex comparisons of other factors in raw LGG data

TableS6 Sex comparisons of other factors in balanced GBM data

Table S7 Sex comparisons of other factors in balanced LGG data

Table S8 Significance of permutation tests for comparison of mutation burden
Attachment

Table S9 Driver genes showing tendency to be clonal in males or females with GBM.
Attachment

Table S10 Driver genes showing tendency to be clonal in males or females with LGG.
Attachment

Supplementary Figure Legends

Figure S1. Overall mutation burden comparison between men and women with GBM or LGG.

Figure S2. Clonal and subclonal mutation burden comparison between men and women. (A) Comparison of clonal and subclonal mutation burden between two gender groups across GBM or LGG. Significance from Wilcoxon rank-sum test is indicated. (B) Comparison of non-silent clonal mutation number and subclonal mutation number between males and females across GBM and LGG. (C) Comparison between males and females across grade 3 and grade 4 gliomas. (D) Comparison between males and females across astrocytoma and glioblastoma. (E) Comparison between males and females in IDH mut and 1p/19q codel subtype.

Figure S3. Subclonal mutation burden comparison between male and female across different chromosomes. (A) Subclonal mutation burden comparison between genders in each chromosome of GBM. (B) Subclonal mutation burden comparison between genders in each chromosome of LGG.

Figure S4. The correlation of subclonal mutation number and confounding factors in patients with GBM or LGG. (A) The correlation of subclonal mutation number and age in patients with LGG excluded X chromosome. (B) The correlation of subclonal

mutation number and tumor purity in patients with GBM or LGG.

Figure S5. Clonal and subclonal mutation burden comparison in balanced GBM and LGG data. (A-D)The distributions of propensity scores in male and female patients of (A) raw GBM data, (B) balanced GBM data, (C) raw LGG data and (D) balanced LGG data. (E-F) Comparison of clonal and subclonal mutation burden between two gender groups across balanced GBM data. (G-H) Comparison of clonal and subclonal mutation burden between two gender groups across balanced LGG data.

Figure S6. The clonal mutation fraction in driver genes versus that of silent mutations and non-driver genes in different gender patients with glioma. Red and blue colors represent clonal and subclonal mutations, respectively.

Figure S7. The cancer cell fraction of mutations in driver genes showing a sex-specific clonal tendency. (A) The cancer cell fraction of mutations in driver genes showing a male-specific tendency to be clonal in GBM. (B) The cancer cell fraction of mutations in driver genes showing a sex-specific clonal tendency in LGG

Figure S8. Kaplan–Meier estimates of OS in three male groups with GBM. Orange curve represents samples carrying *PTEN* clonal mutations. Blue curve represents samples carrying *PTEN* subclonal mutations. Green curve represents samples without *PTEN* mutations.

Figure S9. Prognostic value of *PTEN* clonality in the validation cohort. Survival curves were plot according to clonal status of *PTEN* in female patients (left) and male patients (right).

Figure S10. Effect of X chromosome in mutational burden comparison and the distribution of mutation clonal status for driver genes in each subtype. Subclonal mutation burden comparison between genders in X chromosome of transcriptome subtypes of GBM (A) and molecular subtypes of LGG (B). When excluding mutations in X chromosome, subclonal mutation burden comparison between genders in transcriptome subtypes of GBM (C) and molecular subtypes of LGG (D). (E) The number of clonal mutation and subclonal mutation of GBM driver genes in four transcriptomic subtypes. P value of clonal fraction comparison between non-silent mutations of driver genes and background silent mutations is indicated above each bar. (F) The number of clonal mutation and subclonal mutation of LGG driver genes in three molecular subtypes.

Figure S11. Comparison of the distribution of variant allele frequencies (VAFs) between indel mutations and clonal single nucleotide variants (SNVs). (A) The density of the observed VAFs for each indel group and SNVs.*BF* represent the estimated bias factor. (B) The number of indel mutations in each length group.

Table S1 Detailed patient statistics across GBM and LGG cohorts.

	GBM			LGG			
	Total (N=590)	IDH Mutation (N=17)	IDH Wild Type (N=278)	Total (N=515)	IDH Mutation and 1p/19q Codeletion (N=169)	IDH Mutation and No 1p/19q Codeletion (N=248)	IDH Wild Type (N=91)
Age median	59	40	63	41	45	36	57
interval	10-89	21-66	21-89	14-87	17-75	14-73	21-87
⩽60 years	324	16	124	400	130	207	57
> 60years	266	1	154	57	21	10	26
Gender							
male	362	13	174	285	93	138	51
female	228	4	104	230	76	110	40
Histology							
oligodendroglioma	0	0	0	174	117	36	18
astrocytoma	0	0	0	169	4	112	51
glioblastoma	590	17	278	0	0	0	0
oligoastrocytoma, NOS#	0	0	0	61	17	42	2
anaplastic oligoastrocytoma, NOS#	0	0	0	53	13	27	12
Grade							
G2	0	0	0	249	94	131	20
G3	0	0	0	266	75	117	71
G4	590	17	278	0	0	0	0
Survival data	589	17	277	457	151	217	83
Mutation data	295	17	278	508	169	248	91
SNP6.0 data	577	16	270	513	169	246	90

Since the diagnosis of oligoastrocytoma is strongly discouraged in the 4th revised version of WHO guidelines, oligoastrocytomas were not included as a histologic subtype when conducting analysis.

Table S4 Sex comparisons of other factors in raw GBM data

Variable	Female		Male		Comparisons
	Mean	SD	Mean	SD	Standardized Difference
	N=92		N=161		
Propensity score	0.39	0.085	0.350.11		47%*
Age	61.8	13.6	61.1	12.7	5.7%
Tumor purity	0.76	0.13	0.71	0.16	34%*
Race (yes/no):					
Asian	0	0	0.031	0.17	-25%*
Black or African American	0.076	0.27	0.056	0.23	8.1%
white	0.92	0.27	0.91	0.28	4.0%
IDH status (mut/wt)	0.033	0.18	0.068	0.25	-16.3%*

Table S5 Sex comparisons of other factors in raw LGG data

Variable	Female		Male		Comparisons
	Mean	SD	Mean	SD	Standardized Difference
	N=121		N=170		
Propensity score	0.043	0.074	0.041	0.068	29.8%*
Age	45.4	13.6	43.6	13.8	13.7%*
Tumor purity	0.74	0.15	0.75	0.17	-10.5%*
Race (yes/no):					
American Indian	0	0	0.0058	0.077	-10.8%*
Asian	0.017	0.13	0.0058	0.077	10.1%*
Black or African American	0.049	0.22	0.035	0.19	7.1%
white	0.930.25		0.950.21		-8.2%
IDH status (mut/wt)	0.810.39		0.81 0.39		-0.0047%
1p/19qcodel (yes/no)	0.37	0.49	0.34	0.47	7.6%
Histology(yes/no):					
astrocytoma	0.35	0.48	0.31	0.46	7.5%
oligoastrocytoma	0.29	0.46	0.30	0.46	-2.3%
oligodendroglioma	0.36	0.46	0.39	0.49	-5.2%
WHO Grade(G3/G2)	0.56	0.49	0.53	0.50	6.5%

TableS6 Sex comparisons of other factors in balanced GBM data

Variable	Female		Male		Comparisons
	Mean	SD	Mean	SD	Standardized Difference
	N=92		N=92		
Propensity score	0.39	0.085	0.39	0.082	3.3%
Age	61.8	13.6	62.3	11.8	-3.6%
Tumor purity	0.76	0.13	0.76	0.14	3.9%
Race (yes/no):					
Asian	0	0	0	0	NA
Black or African American	0.076	0.27	0.087	0.28	-4.0%
white	0.92	0.27	0.91	0.28	4.0%
IDH status (mut/wt)	0.033	0.18	0.033	0.18	0

Table S7 Sex comparisons of other factors in balanced LGG data

Variable	Female		Male		Comparisons
	Mean	SD	Mean	SD	Standardized Difference
	N=121		N=121		
Propensity score	0.043	0.074	0.042	0.062	9.7%
Age	45.4	13.6	45.6	14.1	-0.0089%
Tumor purity	0.74	0.15	0.74	0.18	-0.035%
Race (yes/no):					
American Indian	0	0	0	0	NA
Asian	0.017	0.13	0.0082	0.09	7.4%
Black or African American	0.049	0.22	0.049	0.22	0
white	0.93	0.25	0.94	0.23	-3.4%
IDH status (mut/wt)	0.81	0.39	0.79	0.41	0.041%
1p/19qcodel (yes/no)	0.37	0.49	0.34	0.48	6.9%
Histology(yes/no):					
astrocytoma	0.35	0.48	0.35	0.48	0
oligoastrocytoma	0.29	0.46	0.29	0.46	0
oligodendroglioma	0.36	0.46	0.36	0.46	0
WHO Grade(G3/G2)	0.56	0.49	0.55	0.49	7.0%

Figure S1. Overall mutation burden comparison between men and women with GBM or LGG.

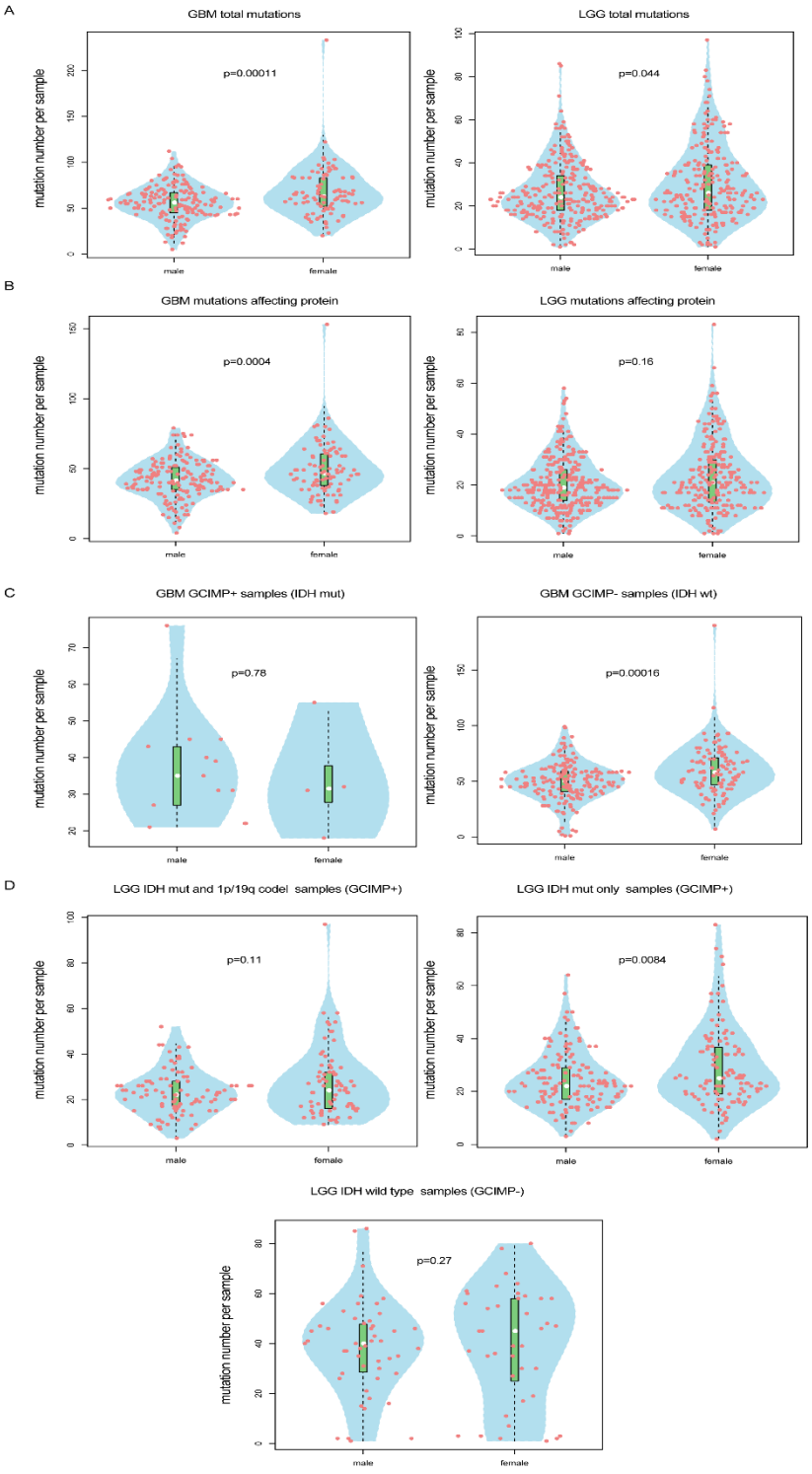


Figure S2. Clonal and subclonal mutation burden comparison between men and women.

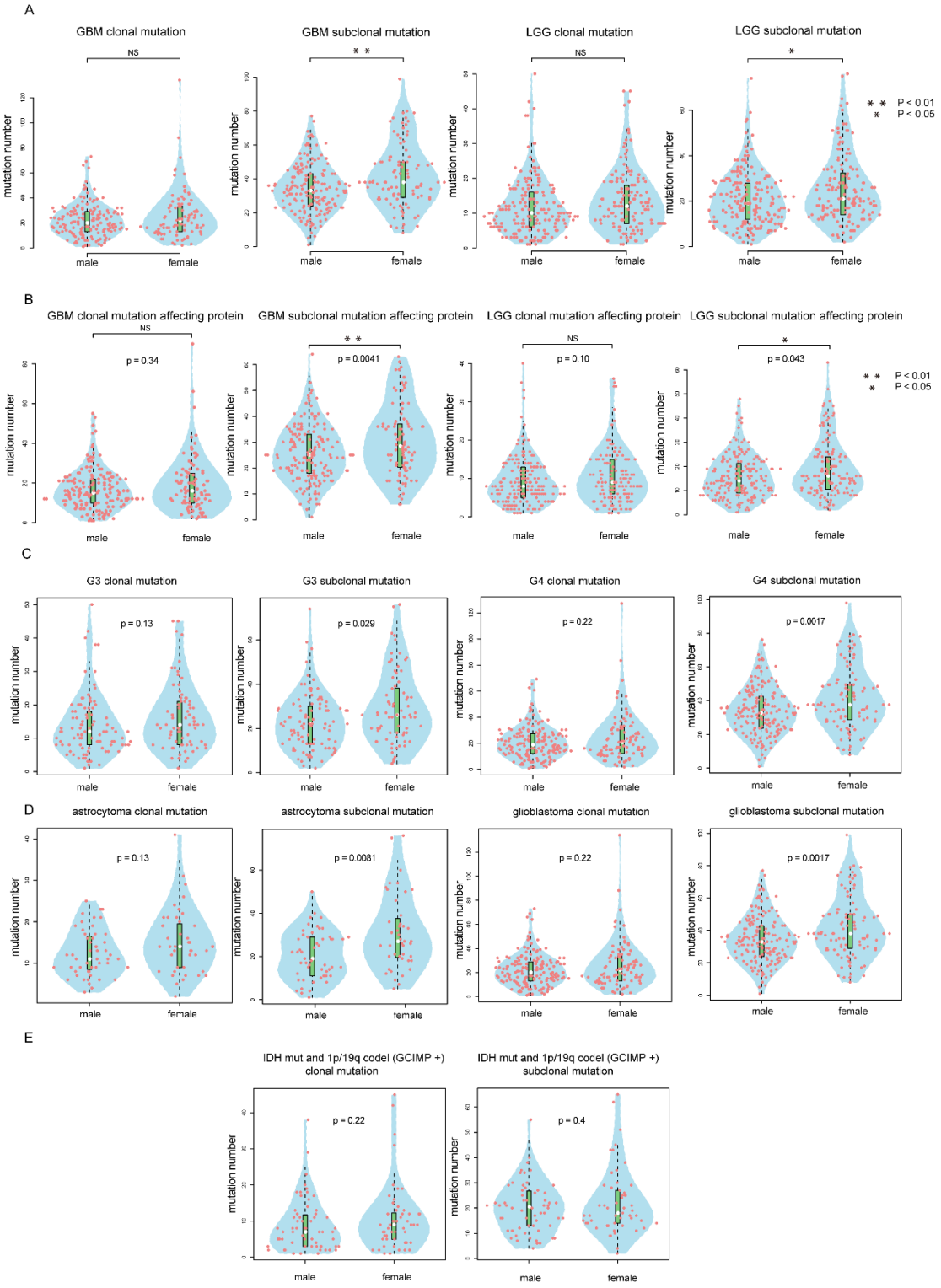
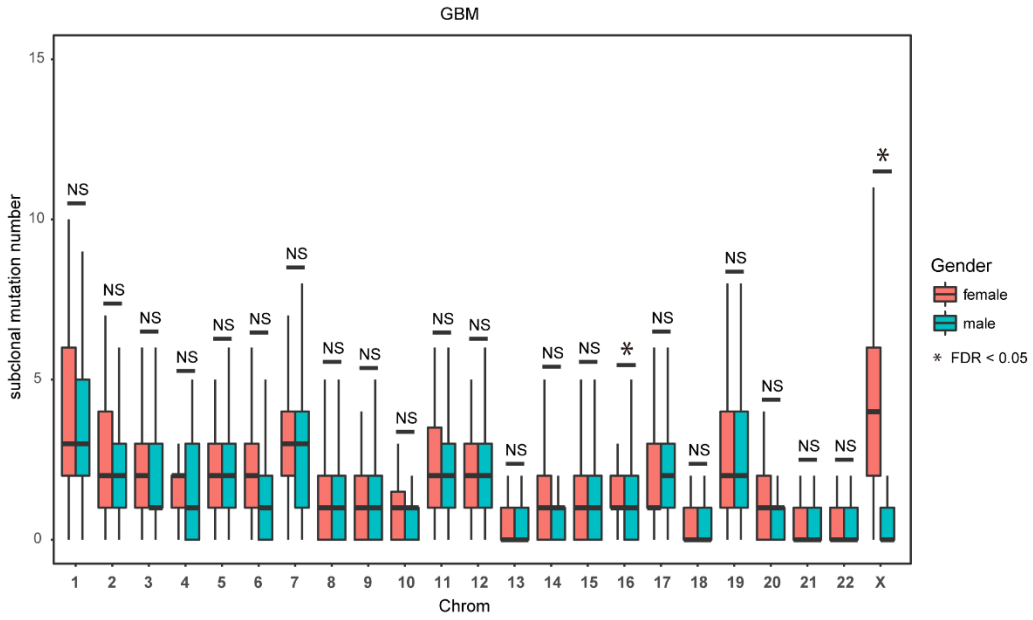


Figure S3. Subclonal mutation burden comparison between male and female across different chromosomes.

A



B

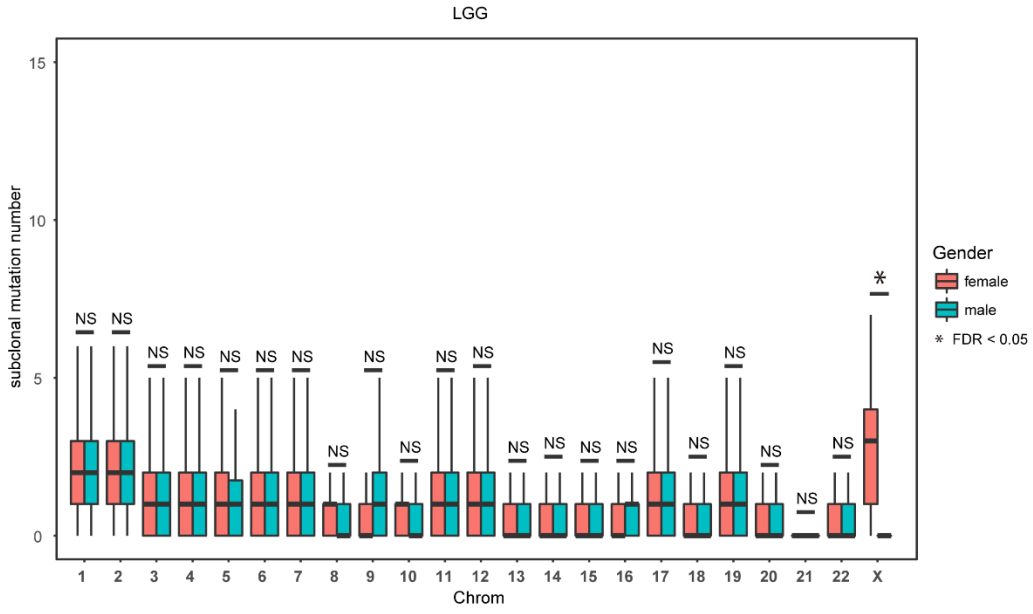
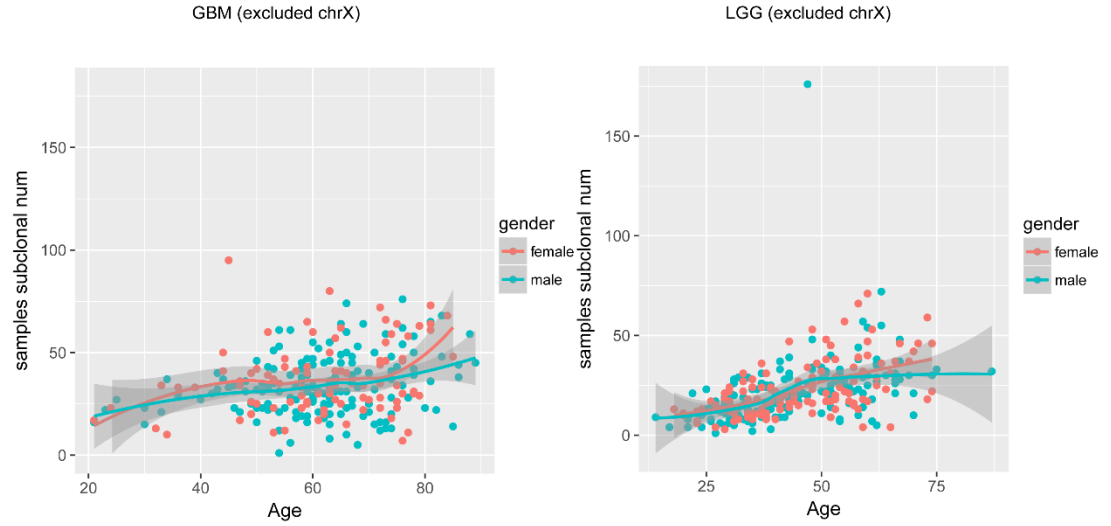


Figure S4. The correlation of subclonal mutation number and confounding factors in patients with GBM or LGG.

A



B

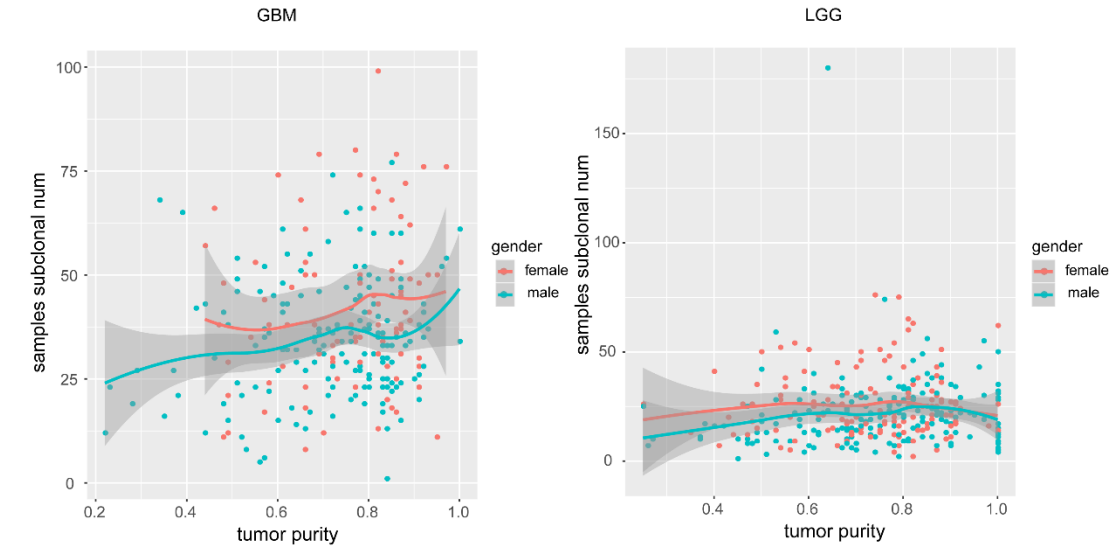


Figure S5. Clonal and subclonal mutation burden comparison in balanced GBM and LGG data

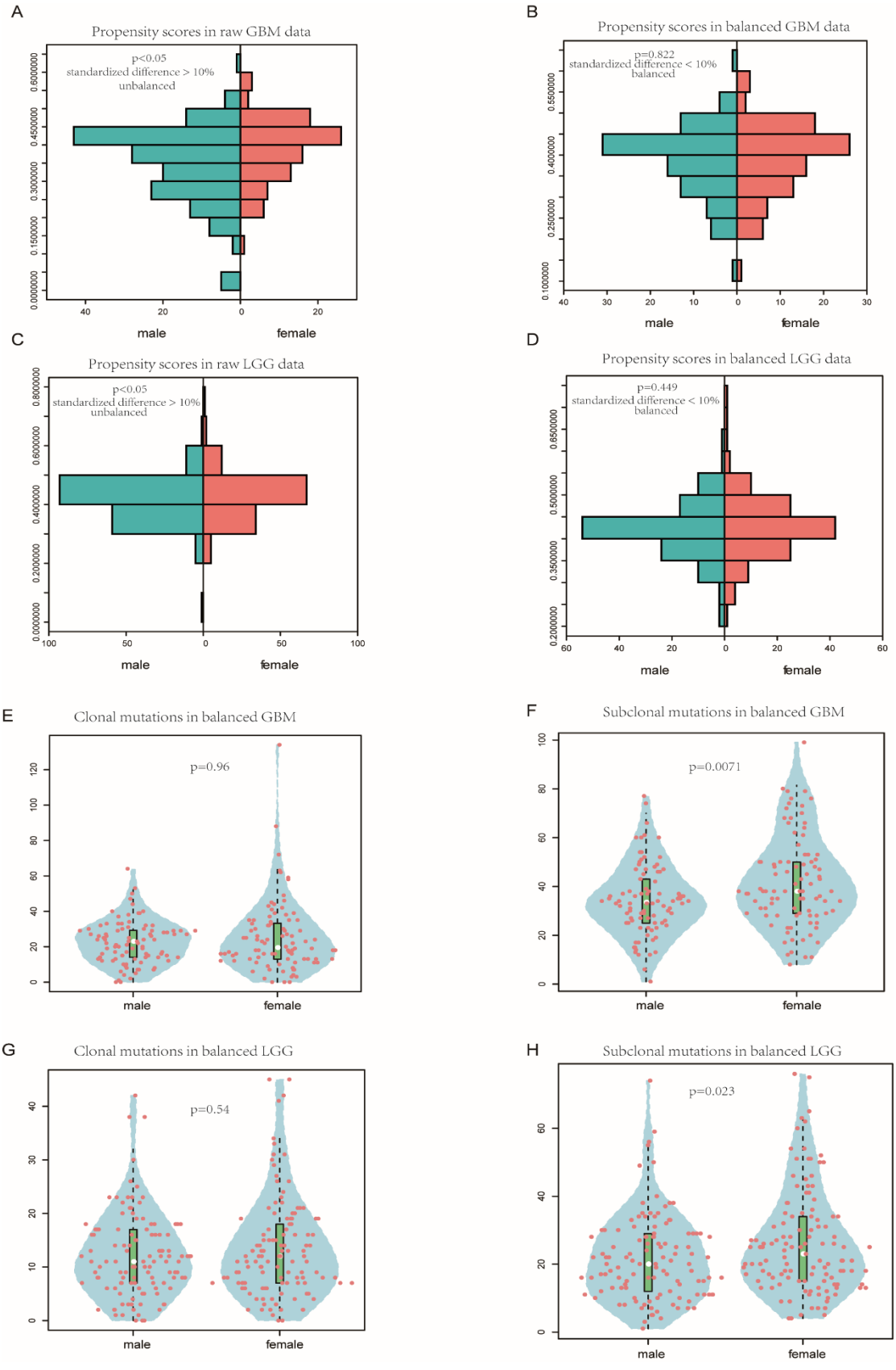


Figure S6. The clonal mutation fraction in driver genes versus that of silent mutations and non-driver genes in different gender patients with glioma.

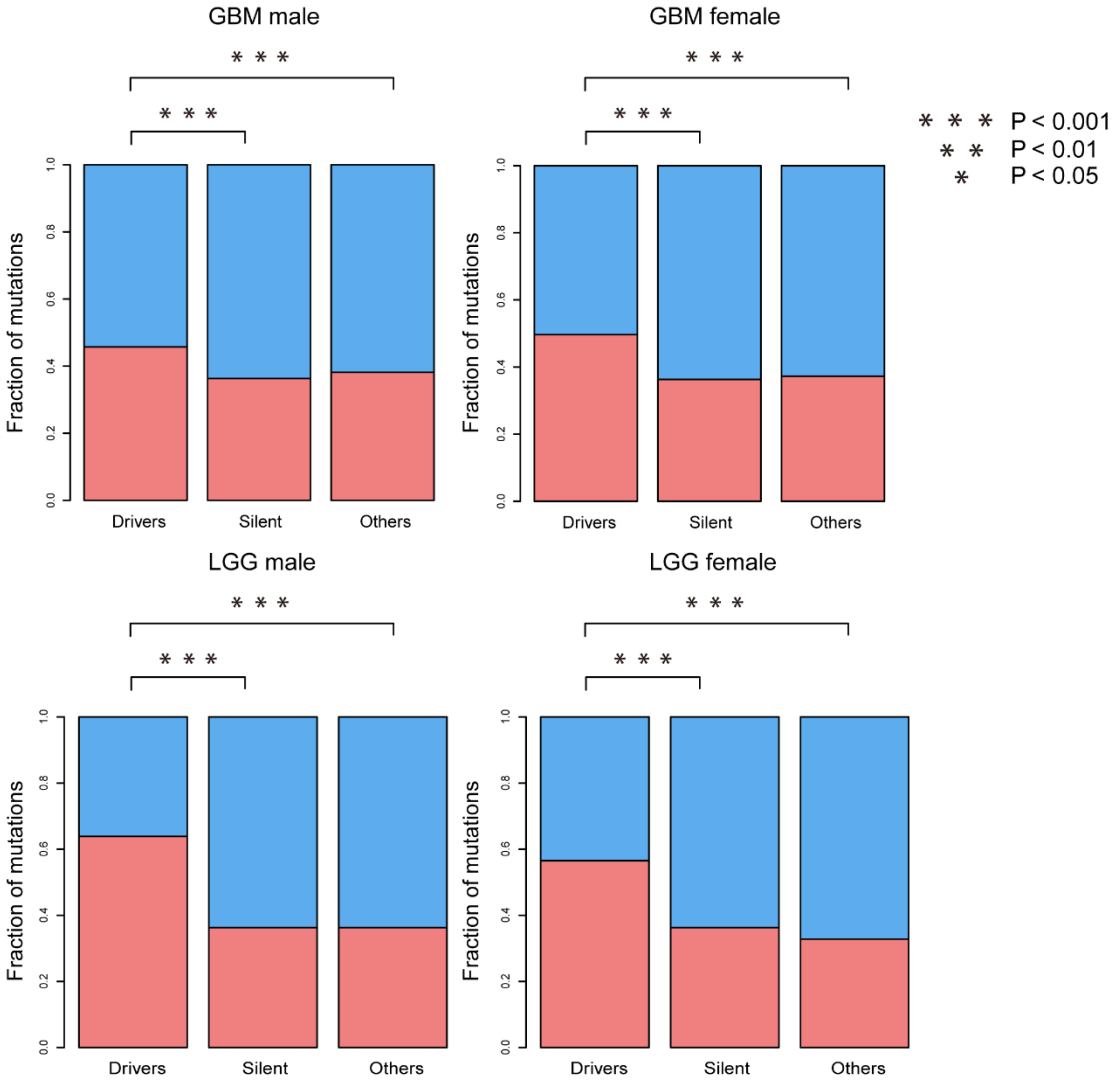


Figure S7. The cancer cell fraction of mutations in driver genes showing a sex-specific clonal tendency.

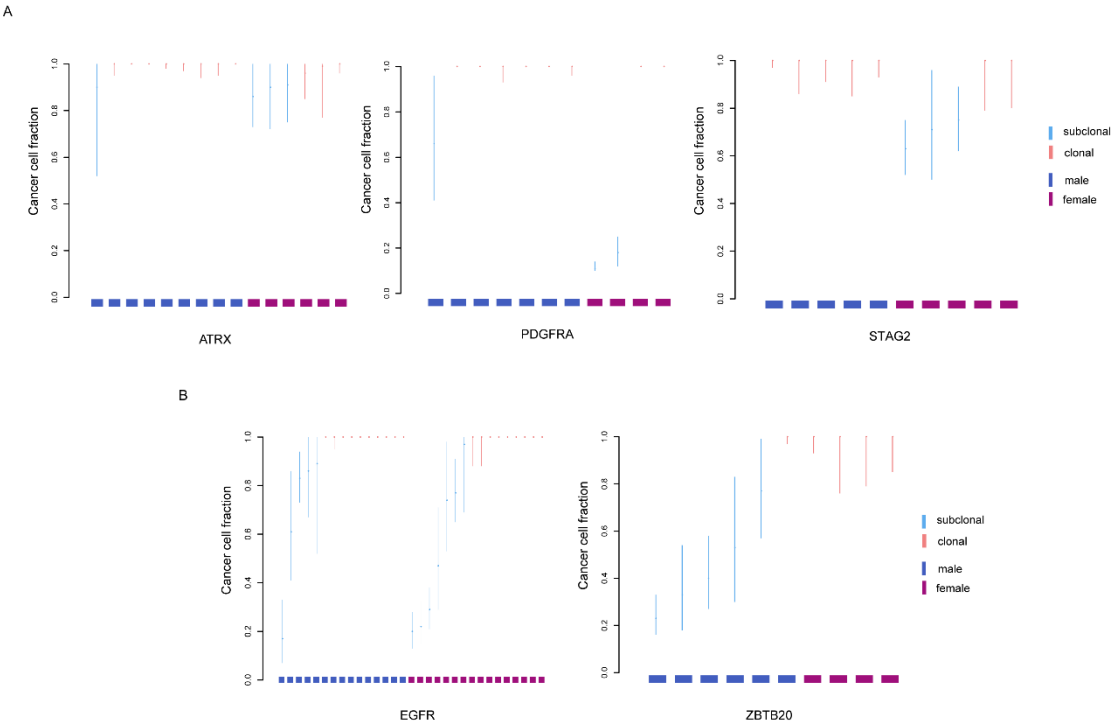


Figure S8. Kaplan–Meier estimates of OS in three male groups with GBM.

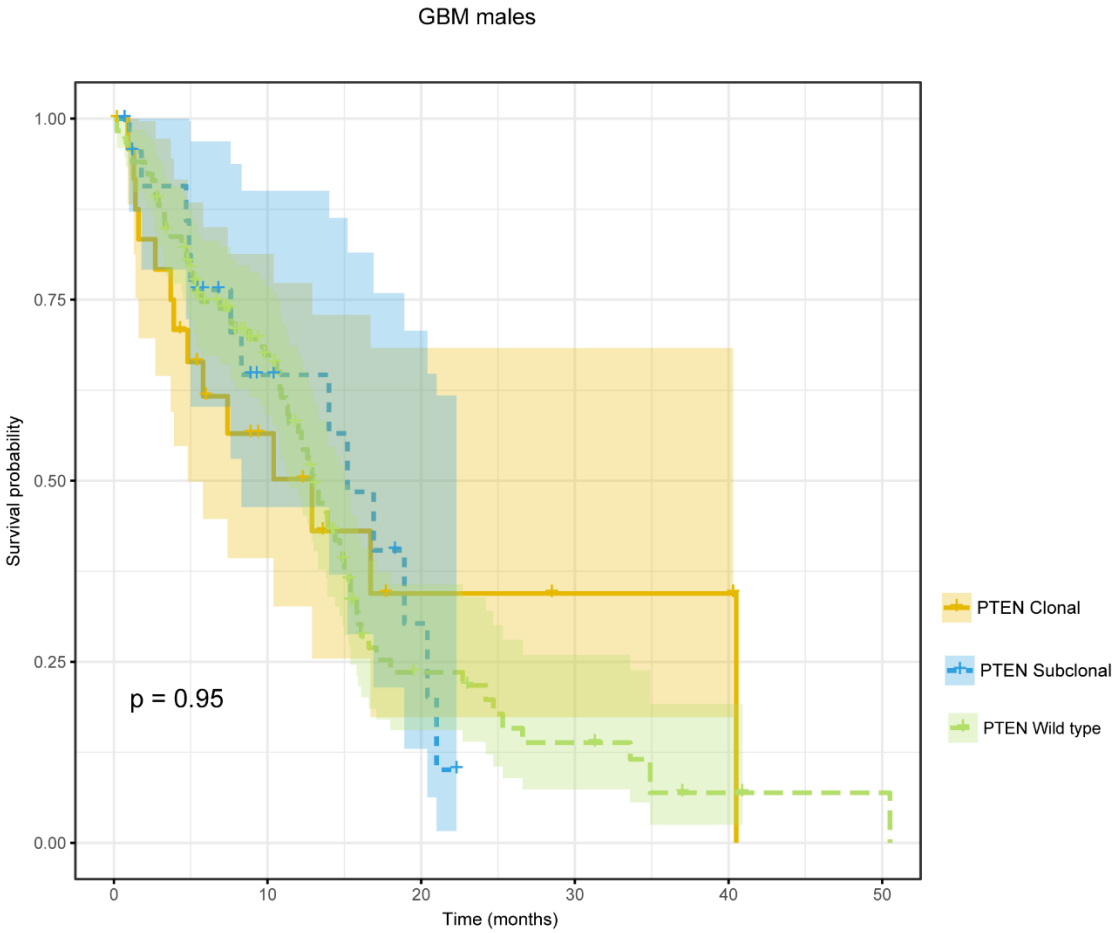


Figure S9. Prognostic value of PTEN clonality in the validation cohort

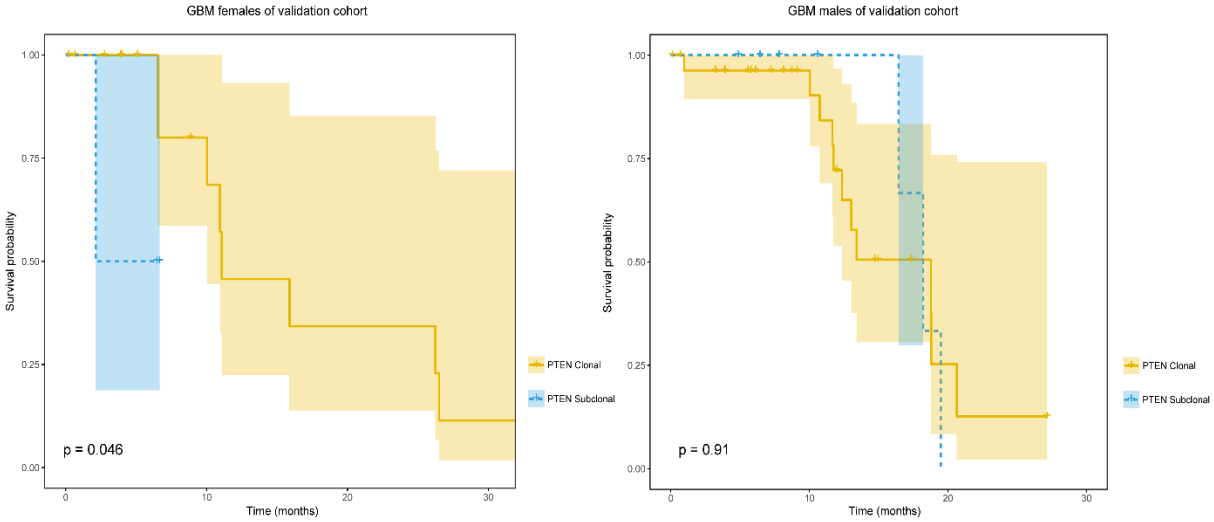


Figure S10. Effect of X chromosome in mutational burden comparison and the distribution of mutation clonal status for driver genes in each subtype.

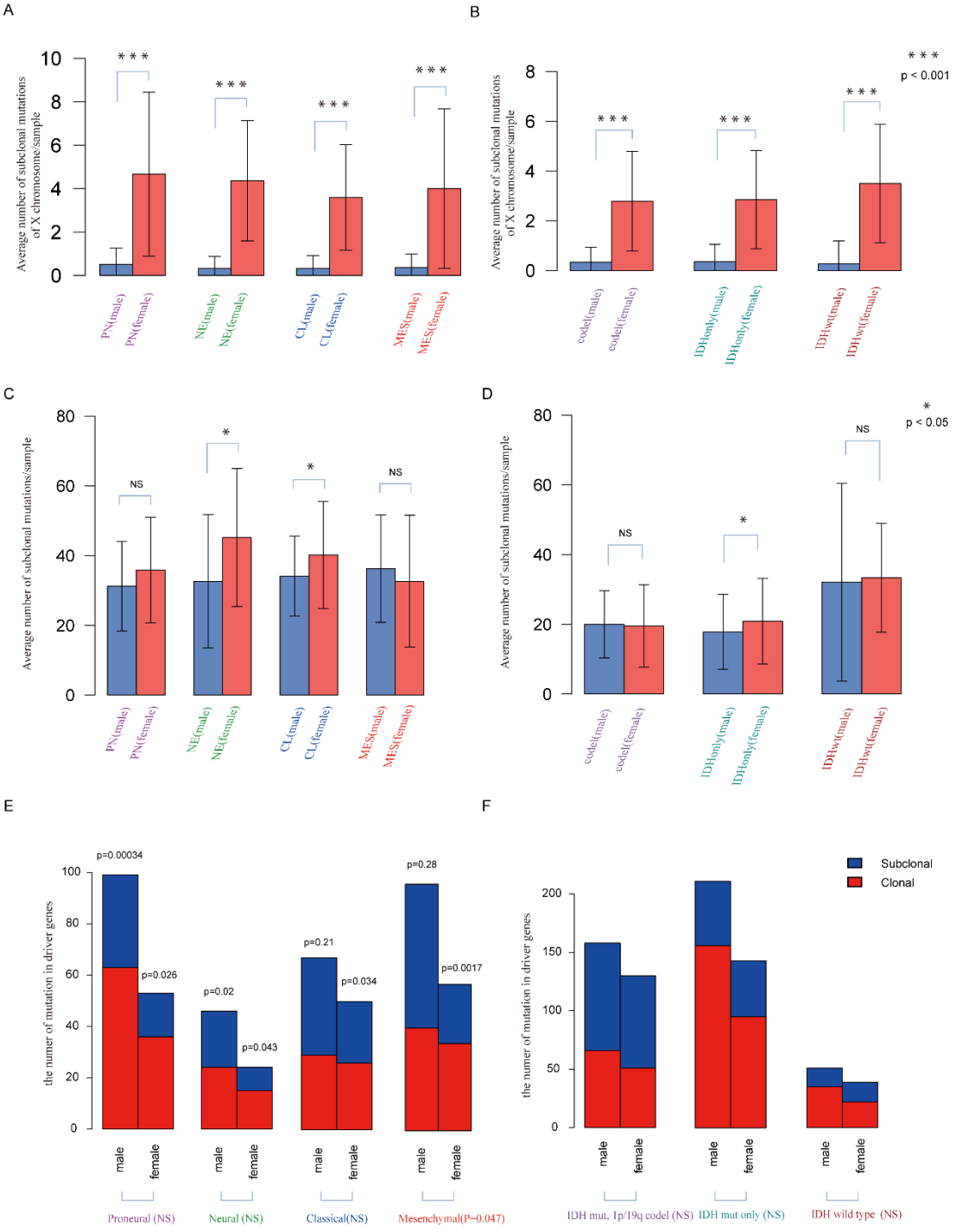


Figure S11. Comparison of the distribution of variant allele frequencies (VAFs) between indel mutations and clonal single nucleotide variants (SNVs).